Data Appendix

Trans Risk Behaviors

March 26, 2018

```
##load data from main rmd file.RMD
##data is filtered_data with all updated categories
##select the labeled versions of the variables
cleardata <- filtered_data %>%
  select(transgender, high_hiv_risk, race, racebinary, generalhealth, physicalhealth, mentalhealth, poo
##Look at the structure of the data
str(cleardata)
```

```
## 'data.frame':    198170 obs. of  26 variables:
##  $ transgender      : chr  "Cis" "Cis" "Cis" "Cis" ...
##  $ high_hiv_risk    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ race             : chr  "white" "white" "hispanic" "white" ...
##  $ racebinary       : chr  "White" "White" "Non-White" "White" ...
##  $ generalhealth    : int  2 4 1 2 2 4 1 1 2 4 ...
##  $ physicalhealth   : num  2 0 7 0 1 1 15 2 0 30 ...
##  $ mentalhealth     : num  2 22 0 0 0 0 0 0 0 30 ...
##  $ poorhealth       : num  2 1 3 NA 0 1 0 0 NA 0 ...
##  $ healthplan       : chr  "Healthcare" "Healthcare" "Healthcare" "Healthcare" ...
##  $ medicalcost      : chr  "No cost Barrier" "No cost Barrier" "No cost Barrier" "No cost Barrier" .
##  $ checkup          : chr  "<1" "<1" "<1" "<1" ...
##  $ state            : Factor w/ 26 levels "California","Connecticut",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ exercise         : chr  "Exercise" "No Exercise" "Exercise" "Exercise" ...
##  $ sleep            : int  8 4 9 8 6 7 8 6 6 9 ...
##  $ sexfactor        : chr  "Female" "Female" "Male" "Female" ...
##  $ educationbinary  : chr  "Did Not Graduate College or Technical School" "Did Not Graduate College
##  $ employment       : chr  "Unpaid Work" "Paid Work" "Not working" "Not working" ...
##  $ numchildren      : num  0 0 0 0 0 0 1 0 0 0 ...
##  $ income           : Factor w/ 8 levels "<10k","10k-15k",..: 7 6 NA 7 8 8 8 8 4 NA ...
##  $ incomebinary     : chr  "Greater than $25,000 per year" "Greater than $25,000 per year" "Greater
##  $ hivtest          : chr  "Tested for HIV" "Not Tested for HIV" "Not Tested for HIV" "Tested for HI
##  $ doctorvisits     : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ medicationcost   : chr  NA NA NA NA ...
##  $ sexualorientation: chr  "Straight" "Straight" "Straight" "Straight" ...
##  $ emotionalsupport : Factor w/ 5 levels "Always","Usually",..: NA NA NA NA NA NA NA NA NA NA ...
##  $ bmi              : Factor w/ 4 levels "Underweight",..: 3 2 2 4 2 2 4 2 4 2 ...
```

High HIV Risk

```
tally(~high_hiv_risk, data=cleardata)
```

```
## high_hiv_risk
##      0      1
## 190920   7250
```

We are treating HIV risk as a binary, categorical variable. Since we are using this as our response variable, we filtered out any null values. 0 represents an individual without a high HIV risk, and 1 represents an individual with a high HIV risk.

Transgender

```
tally(~transgender, data=cleardata)
```

```
## transgender
##            Cis         female           male non-conforming
##         197405            346            246            173
```

We are treating transgender status as a categorical variable. Since this is our primary explanatory variable, we are keeping it as a categorical variable, and also filtered out any null values.

Race

```
tally(~racebinary, data=cleardata)
```

```
## racebinary
## Non-White     White
##     42056    156114
```

We collapsed race into a binary variable, white and non-white. The magnitude of each level makes sense based on the general makeup of the US. There are no null values which is concerning, because all "Don't know" or "Refused" were put into NA. Further investigation is necessary to make sure the null values were not added to one of the categories.

Socioeconomic status

```
tally(~incomebinary, data=cleardata)
```

```
## incomebinary
## Greater than $25,000 per year    Less than $25,000 per year
##                        154160                         44010
```

We are also treating socioeconomic status as a binary variable. There is a much larger number of people in the "Greater than $25,000 per year" than the "Less than $25,000 per year" group, but we chose the dividing point not on size, but by the approximate poverty line, so this is an acceptable distribution. There should be NA values, just as in the racebinary variable, so we will need to check that these categories are accurate.

Education

```
tally(~educationbinary, data=cleardata)
```

```
## educationbinary
## Did Not Graduate College or Technical School
##                                       123128
##        Graduated College or Technical School
##                                        74641
##                                         <NA>
##                                          401
```

There are many NA values for this variable, because "Refused" and "Don't Know/Unsure" categories were included in NA. It will be important to keep this in mind when doing our analysis as the reason people answered in one of those two categories might be relevant, but since that is unlikely, we will leave the data as

is for now. All participants were asked this question, so the null values are because of interviewee response, not interviewer choice.

Emotional Support

```
tally(~emotionalsupport, data=cleardata)
```

```
## emotionalsupport
##    Always   Usually Sometimes    Rarely     Never      <NA>
##     13612      5584      2407       674      1016    174877
```

There are many NA values for this variable. This is because most states did not choose to ask this question. Because of this we will have to be careful about how general we make our conclusion with this data. It is possible that the states that asked this question will have different levels of HIV risk than states that didn't, which would confound our data if we filtered out the null values.

Most pressing data cleaning issues

Investigate racebinary and educationbinary for missing null values.

If we decide to use addition variables from above, we will need to repeat this process.

Rename the levels for high_HIV_risk into a clearer version.